A Software Retina for Egocentric & Robotic Vision Applications on Mobile Platforms

J. Paul Siebert[1], Adam Schmidt[2], Gerardo Aragon-Camarasa[1], Nick Hockings[3], Xiaomeng Wang[1] & W. Paul Cockshott[1]

[1]School of Computing Science, Computer Vision for Autonomous systems Group, University of Glasgow, Scotland, UK; [2]CETAPS laboratory, Faculty of Sports Sciences, Rouen University, Rouen, France; [3]Robotics Laboratory, School of Computer Science, University of Bath, Bath, UK.

Abstract. We present work in progress to develop a low-cost highly integrated camera sensor for egocentric and robotic vision. Our underlying approach is to address current limitations to image analysis by Deep Convolutional Neural Networks, such as the requirement to learn simple scale and rotation transformations, which contribute to the large computational demands for training and opaqueness of the learned structure, by applying structural constraints based on known properties of the human visual system. We propose to apply a version of the retino-cortical transform to reduce the dimensionality of the input image space by a factor of ~x100, and map this spatially to transform rotations and scale changes into spatial shifts. By reducing the input image size accordingly, and therefore learning requirements, we aim to develop compact and lightweight egocentric and robot vision sensor using a smartphone as the target platform.

Keywords: egocentric vision, retino-cortical transform, smart cameras, Deep Learning, CNN, robot vision, biologically motivated computer vision

1 Introduction

A key design issue common to both egocentric and robotics vision systems is the cost of an integrated camera sensor that meets the bandwidth/processing requirement for many applications underpinned by advanced visual perception capability. This is especially the case for ego-centric vision applications that are inherently restricted to lightweight processing and also certain robotics applications, such as SLAM or surveillance in autonomous aerial vehicles. There is currently much work going on to adapt smartphones to provide complete portable vision systems, as the relatively low-cost and ubiquitously available smartphone is so exquisitely integrated by having camera(s), inertial sensing, sound inout/output and excellent wireless connectivity. Mass market production makes this a very low-cost platform and manufacturers from quadrotor drone suppliers to childrens toys, such as the Meccanoid robot, employ a smartphone to provide a vision system/control system.

Accordingly, many research groups are attempting to optimise image analysis, computer vision and machine learning libraries for the smartphone platform. However current approaches to ego-centric and robot vision remain highly demanding for mobile processors such as the ARM, and while a number of algorithms have been developed, these are very stripped down, i.e. highly compromised in function or performance. For example, in order to obtain a semi-dense visual odometry implementation on a smartphone this must be constrained to operate on images of only 320x240pixels. In this work, we are considering the above issues in the design of a biologically motivated practical vision system applied to ego-centric- and lightweight robotics systems based on using a smartphone platform.

2 The Eye-brain Mapping

Nature appears to have solved the above issue of resolution-versus-bandwidth by adopting the space variant sampling adopted by the retina. Vision systems based on the foveated architectures found in mammals have also the potential to reduce bandwidth and processing requirements by about x100 - it has been estimated that our brains would weigh ~ 60Kg if we were to process all our visual input at uniformly high resolution. The central 5 degrees of the visual field are sampled with an almost uniform dense tessellation and beyond this the sampling density falls exponentially, transforming into a complex-log spiral tessellation. In effect nature implements a form of zoom lens that provides a very high acuity central visual field, while at the same time affording around 150 degrees of vision (per eye), yet also limiting the number of photoreceptors to manageable numbers.

The above sampling strategy affords a number of interesting properties: the number of visual receptors sampling a contour centred in the visual field remains approximately constant irrespective of the contours size, potentially offering size-resolution invariance. The effect of this size-resolution invariance is of course to reduce the resolution of the image as a function of off-axis distance which results in loss of high frequency peripheral detail. This can be interpreted as an advantage by removing small-scale clutter from the peripheral field and allowing only large scale contextual structures to remain in view. Accordingly the retinal sampling structure acts as an attentional searchlight, focusing visual processing where it is required and in effect serialising visual search by forcing the visual system to redirect its gaze to sample new and salient visual information, while simultaneously controlling the complexity of the visual input that must be processed during any single observation.

The above retinal sampling strategy is accompanied by a spatial transformation that becomes apparent in the Lateral Geniculate Nucleus en-route to visual area 1 (V1). The mapping, also known as the retino-cortical (RC) transform is well known, having been reported for example by Schwartz [1] ,[2], Johnston [3], Boduc [4] and Gomes [5], and essentially attempts to characterise the combined retinal photoreceptor sampling strategy and brain mapping in terms of either simple analytic functions or composite mappings that separate a quasi-

uniform fovea and its surrounding space-variant periphery. The RC transform in its simplest form approximates a complex-log mapping which translates scale and rotation changes into orthogonal shifts in the output image. This has the advantage in the context of ego-centric vision is that a directly viewed optical flow field generated by egomotion will then approximate lateral shifts in the output space, reducing the required search raneg to support inter-frame flow-field matching. While an apparent limitation of this simple model is that it is shift-variant, i.e. input image patterns must remain centred in the field of view for scale and rotation invariance to hold, this is less of an issue in the context of ego-centric image analysis, due to the inherently directed nature of the camera in this scenario. A more serious limitation is that the pure complex-log transform does not model the quasi-uniform sampling density of the fovea, instead producing a (sampling-density) singularity at the centre of the visual field.

3 Approach

Based on the above observations, we developed a biologically motivated foveated vision system based on a model of the mammalian retina [6] that circumvents the singularity disadvantage of the simple complex-log retina model. This foveated visual architecture [7] implements a functional model of the retina-visual cortex based on a pseudo-random hexagonal sampling tessellation, generated by an annealing process originally proposed by Clippingdale & Wilson [8]. Our *software retina* was originally designed to produce feature vectors that can be matched & classified using conventional methods and has been adopted, for example, by Williamson [9], to implement a surrogate for the human visual system for visual task performance evaluation.

In the work reported here we are proposing to investigate coupling the highly reduced output of the software retina as the input to a DCNN processing architecture to perform classification and interpretation, thereby matching the bandwidths of the sensor and DCNN processing systems. Given the inherent processing limitations of mobile computing platforms, a viable way forward would be to perform off-line learning and implement the forward recognition path on the mobile platform, returning simple object labels, or sparse hierarchical feature symbols, and gaze control commands to the host robot vision system and controller.

We are now at the early stages of investigating how best to port our foveated architecture onto an ARM-based smartphone platform. To achieve the required levels of performance we propose to port and optimise our retina model to the mobile ARM processor architecture in conjunction with their integrated GPUs. We will then be in the position to provide a foveated smart vision system on a smartphone with the advantage of processing speed gains and bandwidth optimisations. Our approach will be to develop efficient parallelising compilers and perhaps propose new processor architectural features to support this approach to computer vision, e.g. efficient processing of hexagonally sampled foveated images.

Our current goal is to have a foveated system running in real-time on at least a 1080p input video stream to serve as a front-end robot sensor for tasks such as general purpose object recognition using the commercial off-the-shelf smartphone. Initially this system would communicate a (learned) symbol stream to conventional hardware performing back-end visual classification/interpretation, and simple object detection and recognition tasks should also be possible onboard the device. We propose to mount the smartphone hosting this system on a pan-tilt unit and to drive this initially using the feature-based visual attention mechanism developed in [6] to serve robotics applications. To date we have demonstrated this gaze control mechanism for object appearance learning using a conventional PZT camera in conjunction with the software retina, and thereafter we propose to investigate learning visual saliency directly via DCNNs. For ego-centric vision applications we propose to directly head-mount the smartphone and thereby rely upon the wearer to direct the camera's gaze.

4 Conclusions

In this work we propose that exploiting the functional space-variant architecture of human vision is the key to achieving the necessary data reduction to being able to implement complete visual recognition and learning processes on compute-limited mobile platforms, such as smartphones, in order to achieve low-cost and compact egocentric and robot vision systems.

References

1. Schwartz, E.L.: Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. Biological Cybernetics **25**(4) (1977) 181–194
2. Schwartz, E.L.: Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. Vision Research **20**(8) (1980) 645–669
3. Johnston, A.: The geometry of the topographic map in striate cortex. Vision Research **29**(11) (1989) 1493–1500
4. Bolduc, M., Levine, M.D.: A Real-Time Foveated Sensor with Overlapping Receptive Fields. Real-Time Imaging **3**(3) (1997) 195–212
5. Gomes, H.: Model Learning in Iconic Vision. PhD thesis, University of Edinburgh, School of Informatics, Edinburgh, Scotland, UK (2002)
6. Balasuriya, L.: A computational model of space-variant vision based on a self-organised artificial retina tessellation. PhD thesis, University of Glasgow, Department of Computing Science, Glasgow, Scotland, UK (2006)
7. Balasuriya, L., Siebert, J.: Hierarchical feature extraction using a self-organised retinal receptive field sampling tessellation. Neural Information Processing: Letters and Reviews **10**(4-6) (2006)
8. Clippingdale, S., Wilson, R.: Self-similar Neural Networks Based on a Kohonen Learning Rule. Neural Networks **9**(5) (1996) 747–763
9. Williamson, C., Strachan, E., Siebert, J.: Simulating the effects of laser dazzle on human vision. International Laser Safety Conference 2013 (2013)